Faculty Working Papers

BIMULTIVARIATE REDUNDANCY MAXIMIZATION

Johny K. Johansson and R. Narayan

#138

College of Commerce and Business Administration
University of Illinois at Urbana-Champaign

FACULTY WORKING PAPERS

College of Commerce and Business Administration

University of Illinois at Urbana-Champaign

December 19, 1973

BIMULTIVARIATE REDUNDANCY MAXIMIZATION

Johny K. Johansson and R. Narayan

#138

BIMULTIVARIATE REDUNDANCY MAXIMIZATION[1]

by

Johny K. Johansson and R. Narayan
University of Illinois

Introduction

The relationship between two sets of variables is often analyzed with the help of canonical correlation techniques. The interpretation problems are often severe, however, as soon as more than one pair of variates are significant at the pre-selected level. Stewart and Love (1968) have suggested a method, called "redundancy analysis", to deal with these problems. Miller and Farr (1971) pointed out that the redundancy measure would remain invariant with respect to any orthogonal rotation of the complete set of canonical variates, and that, consequently, canonical correlation was only one special case of a general redundancy analysis.

It can be argued that since the redundancy measure provides a straightforward interpretation of the degree to which two sets of variables covary, one focus of bimultivariate analysis ought to be the maximization of the redundancy contribution from as small a set of variates as possible. As a start in that direction, this paper presents an approach to the maximization of the partial redundancy attributable to the first pair of variates.

---

[1]Several persons contributed to this paper. Jagdish Sheth encouraged us to focus upon the problem and gave valuable feedback throughout; Charles Lewis gave exceedingly helpful assistance on the theory part; and Joseph Kolman and Maurice Tatsuoka contributed many valuable ideas to the testing of the optimizing approach. Funds were made available by the University of Illinois Computer Services Office and the Bureau of Economic and Business Research. The authors want to thank the people involved but also absolve them of the responsibility for any remaining errors.

The Theory

Miller and Farr (1971) show that the redundancy attributable to the first linear combination of the Y variables G_1 is equal to

$$RED_{G_1} = (L_{G_1}/tr(R_{YY}))*(\sum_{i=1}^{I} r^2_{G_1F_i})$$

where L_{G_1} stands for the sum of the loadings of the Y variables upon G_1, tr (R_{YY}) stands for the trace of the correlation matrix of the Y's (equal to the number of criterion variables), and the F_i, i=1,..., I stand for the successive orthogonal factors of the X variables.[2] Because these orthogonal factors together span the space of the X's completely, we have

$$\sum_{i=1}^{I} r^2_{G_1F_1} = R^2_{G_1X} ,$$

so that the redundancy becomes a product of the loadings and the squared multiple correlation:

$$RED_{G_1} = (L_{G_1}/tr(R_{YY}))*R^2_{G_1X}$$

The canonical correlation technique maximizes the latter component of the product, whereas a principal component analysis of the Y's would maximize the loading part. In general, then, neither of these two approaches would maximize the redundancy measure.

---

[2] In what follows, we will treat the Y's as "dependent" and the X's as "independent" -- an inverse relationship is dealt with similarly but yields no new insights so is ignored here. Also, in what follows, the redundancy measure will always refer to the first linear combination of Y's unless otherwise stated. Finally, both the Y's and the X's are assumed standardized.

To derive an expression of the redundancy measure -- our objective function -- in the original variables Y and X we proceed as follows. Let W_{G_1} denote the m by 1 vector of variable weights for the first linear combination G_1. Then we have

$$Y W_{G_1} = G_1 \quad ,$$

with the dimension of the Y matrix equal to n by m, n denoting the number of observations, m the number of Y's. Then for the loadings we have

$$R_{YG_1} = R_{YY} W_{G_1} \quad ,$$

R again denoting the correlation matrix. Since we want the squared loadings we need

$$R_{YG_1}^T R_{YG_1} = W_{G_1}^T R_{YY} R_{YY} W_{G_1} \quad ,$$

the T superscript indicating transpose. We also note for future use that

$$tr(R_{YY}) = m \ .$$

As for the squared multiple correlation, we have first

$$XB = \hat{G}_1 \quad ,$$

as the predicted value of G_1, with B denoting the I by 1 vector of parameter weights. Using a least squares fit, we compute B as

$$B = R_{XX}^{-1} R_{XG_1} \quad .$$

To get a measure of the squared simple correlation between the actual and the predicted G_1s -- which is the squared multiple correlation we are looking for -- we compute first

$$V_{\hat{G}_1 \hat{G}_1} = B^T R_{XX} B$$

$$= B^T R_{XG_1} \quad .$$

where V stands for the variance. Then we get

$$r^2_{\hat{G}_1 \hat{G}_1} = \frac{V_{G_1 G_1}}{V_{G_1 G_1}}$$

$$= \frac{(R^{-1}_{XX} R_{XY} W_{G_1})^T R_{XY} W_{G_1}}{W_{G_1}^T R_{YY} W_{G_1}}$$

$$= \frac{W_{G_1}^T R_{XY}^T R_{XX}^{-1} R_{XY} W_{G_1}}{W_{G_1}^T R_{YY} W_{G_1}}$$

for the correlation between the predicted and actual G's. The complete objective function can then be written

$$RED_{G_1} = \frac{(W_{G_1}^T (R_{YY})^2 W_{G_1})}{m} \frac{(W_{G_1}^T R_{XY}^T R_{XX}^{-1} R_{XY} W_{G_1})}{W_{G_1}^T R_{YY} W_{G_1}} ,$$

which is to be maximized under a normalizing constraint such as $W_{G_1}^T W_{G_1} = 1$, or $W_{G_1}^T R_{YY} W_{G_1} = 1$.

The Algorithm

Since the objective function (1) consists of the product of two quadratic functions, for which a gradient procedure might easily stop at a local maximum, the algorithm employed was a direct search routine (the Hooke-Jeeves algorithm described in detail by Himmelblau, 1972, p. 142).

The basic approach to the maximization routine utilized the fact that the objective function can be written

$$RED = F(a) * H(a,b) \ ,$$

with a,b, denoting the weights of the Y-compound and X-compound, respectively. This is a direct generalization of the function as stated in (1). Then the dynamic programming "knapsack" approach gives a solution as

$$\max_{a,b}\{RED\} = \max_{a}\{F(a) * \max_{b}\{H(a,b)\}\}.$$

That is, for a given vector a, find the vector b that maximizes H(a,b); then, search over feasible vectors a, maximizing H every time, to find the one that maximizes RED. Since for a given a, and thus a given Y-compound, the maximum G is obtained by a multiple regression of the given Y-combination upon all the X variables, the first maximum can be located. Then, considering the constraints, the search can be made over a relatively small number of a-values, namely those that lie within the limits -1.0 to +1.0 for all elements in a.

Thus, the algorithm iterated a search by first picking the trial a's, then getting the loadings of the original Y variables upon the generated linear compound, and finally computing the regression of the compound upon the X variables. The derived R^2, multiplied by the average squared loadings constituted the trial value of the objective

function. A search then generated new a-values, and another iteration took place. The routine would stop iterating when either one of four preset test values was superceded.

The strength of the search routine was abetted by the fact that a generally good starting point could be generated (the canonical correlation weights) and by the fact that the total redundancy between the two sets was given by the average R^2 between each of the Y's and the X's. Thus, the maximum obtainable solution could be checked against it.[3]

The constraint used in the runs was $W_{G_1}^T W_{G_1} = 1$. Initially, each set of trial a's within the [-1, +1] hypercube were scaled so as to fulfill the constraint, before the value of the objective function was computed. This approach impaired the efficiency of the algorithm considerably, however, making it necessary to adopt another approach. The constraint was now tested for, and the a-values scaled, only after the optimal solution had been located. The approximation resulting from this approach was very close to the earlier solution for the problems tested.[4]

Initially, the algorithm was set up for raw data only, but as can be seen from equality (1), the only data input needed would be the correlation matrix of the Y's and the X's. When raw data are input, this correlation matrix is computed at the initial iteration, and the program can then bypass this computation in later iterations.

---

[3]In addition, an alternative search routine, the Nelder-Mead technique of searching successive simplexes (Himmelblau, p. 148), was used for some runs. The optimal solutions located by the two algorithms were the same throughout.

[4]This closeness can be attributed to the fact that the contours of the objective function in all the cases examined formed a ridge in a radial direction from the origin (see Figure 1).

The Results

Initial runs were made on the TALENT data provided by Cooley and Lohnes (1971, Appendix B). The use of published and thus easily accessible data facilitates cross-checks and further analysis. The analysis carried out and their results follow.

The criterion set of variables chosen consisted of "Physical Science Interest", "Office Work Interest", and "Plans to attend College". The predictor set of variables consisted of test scores on "Information test II", "Mechanical Reasoning", and "Reading Ability", plus the student's "Socioeconomic Status". (For further information on the data and these variables, the reader is referred to the Cooley and Lohnes book). The algorithm was run from two starting points, one provided by the criterion weights of a canonical correlation analysis, the other by a principal component analysis of the Y variables. In all cases the search routine isolated the same maximum of the objective function. The runs were made separately for males and females.[5]

The results are presented in Table 1. Overall, they are somewhat surprising in that the redundancy maximization routine only does marginally better than the canonical correlation solution. For the males data the reason is clear: there is very little additional redundancy to account for once the first canonical solution has been taken out. In the female data the reason is less clear -- but one explanation for the almost zero improvement of the redundancy maximization would be that the data are truly explained by two, rather than one, pair of

---

[5]Additional runs were made for males and females combined, as well as for other sets of variables. Since the results were similar to the runs reported here, these other runs are not included.

variates. With these marginal improvements, no great changes are to be expected in the weights -- as can be seen, only minor fluctuations away from the canonical solution occur. The principal components solution, on the other hand, is not as close to the optimal as is the canonical solution. The principal components weights accordingly also show a wider divergence from the redundancy solution.

Since these results were largely reproduced in other runs, it was decided that the objective function be plotted and its behavior more closely examined. As the plotting required one dimension for the function value, plus one additional dimension for each criterion variable, it was decided to plot a case where only two criterion variables were used. Accordingly, the "Office Work Interest" variable was dropped, and the objective function as a function of the ensuing two-element vector a was plotted (the predictor variables remained the same). The values of the resulting objective function for the male data are depicted in Figure 1. As could have been inferred from the earlier runs, the function has a flat ridge around the optimum, making for quite a large near-optimal region.[6] Plots of other runs tended to follow the same pattern. There seems, then, to be a general indication that the canonical solution will quite often be very close to optimizing the redundancy contribution from the first pair of variates.

---

[6]The symmetry of the objective function follows from the fact that a change in sign will not affect the optimal property of the weights. For completeness, the redundancy analysis results for this case with two criterion variables are included in Table 4.

Conclusions and Extensions

Although these initial data runs pointed in the other direction, it is clearly too early to dismiss the possibility that significant changes in the weights -- and hence of the interpretations -- of the original variables can occur when the redundancy attributable to the first pair of variates is maximized rather than its canonical correlation. The theory is unequivocal: the canonical solution will in general not be optimal. In what type of particular data structures it will be approximately optimal remains to be investigated further. One thing seems already quite clear: If only one canonical root is significant at the pre-selected level, chances are that a redundancy maximization will make very little difference.

Although in this paper redundancy was maximized with reference only to the first pair of variates, a straightforward generalization to further variates is easily made. For the optimal linear Y-compound, the loadings of the separate Y-variables are first computed. Using the fundamental factor theorem the amount of variation in the original Y-variables explained by the optimal compound is then derived. The unexplained variation in the Y-variables is what then remains to be explained by a second Y-compound. Similarly, the residual variation in the X-variables after the first X-compound is extracted can be derived. The second redundancy maximization can then take place using the residual variations in the Y and X variables.

List of variables: for TALENT DATA (Males and Females)

Y_1	Plan College full time 1. Definitely will go 2. Almost sure to go 3. Likely to go 4. Not likely to go 5. Definitely will not go
Y_2	Physical Science Interest Inventory
Y_3	Office Work Interest Inventory
X_1	Information Test Part II
X_2	Reading Comprehension Test
X_3	Mechanical Reasoning Test
X_4	Socioeconomic Status Index

TABLE 1

MALES: Total Redundancy = .180

	CC	PC	RED
b_1	-.187	.085	-.161
b_2	-.173	.182	-.184
b_3	-.112	.137	-.124
b_4	-.303	.250	-.299
a_1	.715	-.599	.688
a_2	-.636	.673	-.636
a_3	.292	.435	.312
Redundancy	.171	.136	.179

FEMALES: Total Redundancy = .145

	CC	PC	RED
b_1	.204	.211	.207
b_2	.131	.141	.136
b_3	.144	.101	.129
b_4	.188	.197	.193
a_1	-.690	-.651	-.690
a_2	.651	.515	.609
a_3	-.316	-.558	-.404
Redundancy	.121	.118	.120

TABLE 2

CANONICAL CORRELATIONS

MALE DATA

Function	Eigenvalue	Correlation	Wilks Lambda	Chi-Square	DF
1	0.3727	0.6105	0.6017	116.8313	12
2	0.0339	0.1842	0.9592	9.5756	6
3	0.0071	0.0842	0.9929	1.6373	2

FEMALE DATA

Function	Eigenvalue	Correlation	Wilks Lambda	Chi-Square	DF
1	0.2460	0.4960	0.6851	100.9900	12
2	0.0876	0.2960	0.9086	25.6011	6
3	0.0042	0.0645	0.9958	1.1130	2

TABLE 3

Sample Size = 234

Correlation Matrix — MALE DATA

	X1	X2	X3	X4	Y1	Y2	Y3
	1	2	3	4	5	6	7
1	1.00000						
2	0.79089	1.00000					
3	0.50389	0.44324	1.00000				
4	0.48847	0.36824	0.27985	1.00000			
5	-0.43111	-0.43956	-0.28882	-0.44597	1.00000		
6	0.43540	0.36442	0.34489	0.36142	-0.47276	1.00000	
7	-0.04110	0.00452	0.03433	-0.01529	-0.11191	0.29653	1.00000

	MEANS	STANDARD DEVIATION
[illegible]	[illegible]D 02	0.17882D 02
2	0.33585D 02	0.95132D 01
3	0.13568D 02	0.35819D 01
4	0.98543D 02	0.94442D 01
5	0.29017D 01	0.15596D 01
6	0.21321D 02	0.92181D 01
7	0.12291D 02	0.79013D 01

TABLE 3 (con't)

Sample Size = 271

Correlation Matrix FEMALE DATA

	X1	X2	X3	X4	Y1	Y2	Y3
	1	2	3	4	5	6	7
1	1.00000						
2	0.66609	1.00000					
3	0.50726	0.57799	1.00000				
4	0.30383	0.23813	0.16453	1.00000			
5	-0.32493	-0.33711	-0.19538	-0.32406	1.00000		
6	0.31734	0.27348	0.39062	0.13056	-0.28241	1.00000	
7	-0.24229	-0.20473	-0.11889	-0.18551	0.32471	-0.13402	1.00000

	MEANS	STANDARD DEVIATION
1	0.73786D 02	0.17970D 02
2	0.33860D 02	0.89382D 01
3	0.90627D 01	0.34981D 01
4	0.98531D 02	0.10733D 02
5	0.32362D 01	0.16963D 01
6	0.12066D 02	0.76948D 01
7	0.25317D 02	0.97818D 01

TABLE 4

Males (Two Criterion Variables): Total Redundancy = .2581

	CC	PC	RED
b_1	.133	-.152	.140
b_2	.203	-.183	.196
b_3	.121	-.130	.124
b_4	.302	-.293	.299
a_1	-.808	.707	-.774
a_2	.590	-.707	.634
Redundancy	.2568	.2562	.2574

CANONICAL CORRELATIONS:

Function	Eigenvalue	Correlation	Wilks Lambda	Chi-Square	DF
1	.3507	.5922	.6305	106.32	8
2	.0289	.1700	.9711	6.76	3

	-1.3	-1.1	-.9	-.7	-.5	-.3	-.1	.1	.3	.5	.7	.9	1.1	1.3
1.3	256	253	246	236	220	196	164	125	[illegible]	049	025	012	008	009
1.1	257	256	252	243	228	203	167	122	074	037	162	008	008	012
.9	256	257	256	250	237	213	172	115	060	023	009	008	013	022
.7	251	255	257	256	248	225	179	108	041	011	008	006	028	042
.5	241	246	252	256	256	241	190	093	019	008	020	039	059	070
.3	222	228	235	244	253	256	213	060	008	033	064	087	103	115
.1	195	198	202	208	218	235	256	008	087	124	140	149	154	158
-.1	158	154	149	140	124	087	008	256	235	218	208	202	198	195
-.3	115	103	087	064	033	008	060	213	256	253	244	235	228	222
-.5	074	059	039	020	008	019	093	190	241	256	256	252	246	241
-.7	042	028	016	008	011	041	108	179	225	248	256	257	255	251
-.9	022	013	008	009	023	060	116	172	213	237	250	256	257	256
-1.1	012	008	008	162	037	074	122	167	203	228	243	252	256	257
-1.3	008	008	012	025	049	085	125	164	196	220	236	246	253	256

Y_1

FIGURE I

Plot of the Values of the Objective Function for the Male Data With Two Criterion Variables (Function Values x 10^3).

References

Cooley, W. W. and P. R. Lohnes, Multivariate Data Analysis, New York: Wiley, 1971.

Himmelblau, D. M., Applied Non-Linear Programming, New York: McGraw-Hill, 1972.

Miller, J. K. and S. D. Farr, "Bimultivariate Redundancy: A Comprehensive Measure of Interbattery Relationship", Multivariate Behavioral Research, July 1971.

Stewart, D. and W. Love, "A General Canonical Correlation Index", Psychological Bulletin, September, 1968.